Replication: An Alternative Softmax Operator for Reinforcement Learning

Author names
Department of Computer Science
National Yang Ming Chiao Tung University
{xxx, yyy, zzz}@nycu.edu.tw

1 Problem Overview

Please provide a brief overview of the selected paper. You may want to discuss the following aspects:

- The main research problem tackled by the paper
- High-level description of the proposed method

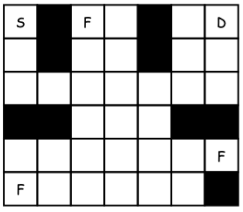
2 Background and The Algorithm

Please present the essential background knowledge and the algorithm in this section. You may also describe the notations and the optimization problem of interest.

3 Detailed Implementation

Please explain your implementation in detail. You may do this with the help of pseudo code or a figure of system architecture. Please also highlight which parts of the algorithm lead to the most difficulty in your implementation.

taxi domain:



environment: In this experience, s is the starting position, d is the destination, and f represents passenger, it will get +1 reward for delivering one passenger, +3 for two , +15 for three.

we preform SARSA with epsilon-greedy, SARSA with Boltzmann softmax and SARSA with Mellowmax softmax. In this environment, there are 33 posible positions and three passengers, therefore, it has 33*2*2*2 = 264 states and four possible actions.

In our algorithm, we set some prohibited state-action pair's Q value to negative infinity, like hit the wall. Moreover, durning the training period, we will not choose those action in order not to affect other Q values.

Because this environment need to have proper exploration, therefore, Boltzmann and Mellowmax can show their advantage. However, there are some detail that not be mention in this paper, so we have our own setting.

In epsilon-greedy method, training and testing both use epsilon-greedy to select action. And in Boltzmann softmax, training and test both use Blotzmann too. However, in Mellowmax softmax, we use Mellowmax softmax durning training process, and in testing process, we select the actions which have max Q values according to our pre-train Q values. We found that there are more approximate to paper's results.

4 Empirical Evaluation

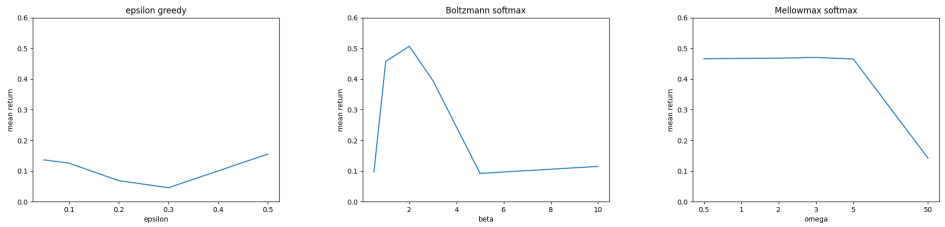
Please showcase your empirical results in this section. Please clearly specify which sets of experiments of the original paper are considered in your report. Please also report the corresponding hyperparameters of each experiment.

taxi domain:

Sarsa with epsilon-greedy, learning rate $\alpha = 0.1$, discount factor $\gamma = 1$, we test different ϵ values, $\epsilon = 0.05, 0.1, 0.2, 0.3, 0.5$, and the result are averaged over 6 independent runs, each consisting of 100000 time steps.

Sarsa with Boltzmann softmax, learning rate $\alpha = 0.1$, discount factor $\gamma = 1$, we test different β values, $\beta = 0.5, 1, 2, 3, 5, 10$, and the result are averaged over 6 independent runs, each consisting of 100000 time steps.

Sarsa with Mellowmax softmax, learning rate $\alpha = 0.1$, discount factor $\gamma = 1$, we test different ω values, $\omega = 0.5, 1, 2, 3, 5, 50$, and the result are averaged over 6 independent runs, each consisting of 10000 time steps. We know that when $\omega \to \infty$, Mellowmax will behave like max operator, and we set ω to 50 so that the result will similar to the paper and show that phenomenon.



(a) ϵ-greedy (b) Boltzmann softmax (c) Mellowmax softmax

Figure 1: Taxi domain result

5 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
- Any technical limitations
- Any latest results on the problem of interest

References